Stats315a Problem Set 2

Due Date: Wednesday, March 6, 11:59pm

Question 2.1 (Resampling methods and normal inference, 40 points)**:** In this problem, we compare three parameter inference methods: (i) model-based inference, which assumes the model is true, (ii) the "sandwich estimator" we derived in class, and (iii) the bootstrap resampling estimator of variance. We will do this both for linear and (binary) logistic regression, repeating the following experimental protocol many times and providing summary results. We first describe the protocol for the Gaussian linear model case; we then describe modifications for the other cases.

i. For the data model

$$y_i = x_i^T \beta^\star + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} \mathsf{N}(0, \sigma^2),$$

generate a sample of size n (to be specified) in dimension $d = 10$, where $x_i \overset{\text{iid}}{\sim} \mathsf{N}(0, I_d)$ and $\sigma^2 = 1$, and draw $\beta^\star \sim \mathsf{Uni}(\mathbb{S}^{d-1})$, the sphere in $\mathbb{R}^d$.

ii. Compute $\widehat{\beta}_n$ minimizing

$$L_n(\beta) = \frac{1}{n} \sum_{i=1}^n (y_i - x_i^T \beta)^2.$$

iii. Consider the following three covariance estimates:

$$\Sigma_n := \widehat{\sigma}^2 (X^T X)^{-1}, \quad \widehat{\sigma}^2 = \frac{1}{n-d} \sum_{i=1}^n (y_i - x_i^T \widehat{\beta}_n)^2 \tag{Fisher}$$

$$\Sigma_n := (X^T X)^{-1} \left(\sum_{i=1}^n (y_i - x_i^T \widehat{\beta}_n)^2 x_i x_i^T \right) (X^T X)^{-1} \tag{Sandwich}$$

$$\Sigma_n := \frac{1}{B} \sum_{b=1}^B (\widehat{\beta}_n^b - \widehat{\beta}_n)(\widehat{\beta}_n^b - \widehat{\beta}_n)^T \tag{Bootstrap}$$

where $\widehat{\beta}_n^b$ is a bootstrap resampled least-squares estimate, and $B = 200$ is the number of bootstrap replicates.

The first covariance is the classical Fisher information matrix, the second the covariance as per the standard asymptotic theory we have developed, and the third that of the boostrap. We have $\widehat{\beta}_n \mathrel{\dot\sim} \mathsf{N}(\beta^\star, \Sigma_n)$ for each of these (so long as the data model remains true!), and therefore

$$\mathcal{C}_n := \left\{ \beta \in \mathbb{R}^d \mid (\beta - \widehat{\beta}_n)^T \Sigma_n^{-1} (\beta - \widehat{\beta}_n) \le \chi^2_{d,1-\alpha} \right\}$$

as an asymptotically valid $1 - \alpha$ confidence set, that is, $\mathbb{P}(\beta^\star \in \mathcal{C}_n) \to 1 - \alpha$, where $\chi^2_{d,1-\alpha}$ is the $1 - \alpha$ quantile of a χ^2 random variable with d degrees of freedom. Use $\alpha = .1$ for the remainder.

(a) Repeat the experiment i–iii for $n = 50, 100, 200, 400$ for $T = 200$ times, and track the fraction of times that $\beta^\star \in \mathcal{C}_n$ for each of the covariances (Fisher), (Sandwich), and (Bootstrap). For each covariance approximation, plot your coverage against sample size n.

(b) Repeat part (a) except for logistic regression. Thus, make the following changes to the procedure i–iii. Instead of the linear regression model, generate data from the logistic regression model

$$\mathbb{P}_\beta(Y = y \mid X = x) = \frac{e^{y\beta^T x}}{1 + e^{x^T\beta}}, \quad y \in \{0, 1\},$$

where $x_i, \beta^\star$ are generated identically. In part ii, choose $\widehat{\beta}_n$ to minimize the negative log likelihood, that is, for $\ell(\beta, x, y) = -\log p_\beta(y \mid x) = \log(1 + e^{x^T\beta}) - yx^T\beta$, let $L_n(\beta) = \frac{1}{n}\sum_{i=1}^n \ell(\beta, x_i, y_i)$. In part iii, replace the Fisher information and Sandwich covariances with their counterparts

$$\Sigma_n := (X^T W X)^{-1}, \quad W = \operatorname{diag}([\widehat{p}_i(1 - \widehat{p}_i)]_{i=1}^n)$$
$$\Sigma_n := \frac{1}{n}\nabla^2 L_n(\widehat{\beta}_n)^{-1}\widehat{\operatorname{Cov}}(\nabla\ell(\widehat{\beta}_n(X, Y)))\nabla^2 L_n(\widehat{\beta}_n)^{-1},$$

respectively (the bootstrap covariance does not change).

(c) Repeat part (a) with faulty linear modeling assumptions, so that we evaluate coverage of $\widehat{\beta}_n$ and $\mathcal{C}_n$ for the best linear predictor in mean-squared error, $\beta^{\text{mse}} = \operatorname{argmin}_\beta \mathbb{E}[(y - x^T\beta)^2]$. To do so, replace the model in part i with

$$y_i = x_i^T\beta^\star + (x_i^T\theta^\star)^2 + \varepsilon_i, \quad \varepsilon_i \stackrel{\text{iid}}{\sim} \mathsf{N}(0, \sigma^2),$$

where $\theta^\star \sim \mathsf{Uni}(\mathbb{S}^{d-1})$ as well. Note that

$$\begin{aligned}\mathbb{E}[(y - x^T\beta)^2] &= \mathbb{E}[(\varepsilon + x^T(\beta - \beta^\star) + (x^T\theta^\star)^2)^2] \\ &= \sigma^2 + \|\beta - \beta^\star\|_2^2 + \mathbb{E}[(x^T\theta^\star)^4],\end{aligned}$$

because $\mathbb{E}[v^T x(u^T x)^2] = 0$ for any vectors u, v,[1] so $\beta^{\text{mse}} = \operatorname{argmin}_\beta \mathbb{E}[(y - x^T\beta)^2] = \beta^\star$ as well.

Question 2.2 (Asymptotics of causal inference, 15 points)**:** As in Question 1.6, consider the potential outcomes framework for a real-valued response Y with randomized treatment assignments $W \in \{0, 1\}$, so that $(Y(0), Y(1)) \perp W$. Let the population mean-square-error estimates be

$$(\tau^{\text{mse}}, \alpha^{\text{mse}}, \beta^{\text{mse}}) = \operatorname*{argmin}_{\tau,\alpha,\beta} \mathbb{E}\left[(Y - \alpha - X^T\beta - \tau W)^2\right],$$

so that $\alpha^{\text{mse}} \in \mathbb{R}$ is an intercept, $\beta^{\text{mse}} \in \mathbb{R}^p$, and τ^{mse} is the coefficient of W in the model $Y_i = \alpha + X_i^T\beta + \tau W_i + \varepsilon_i$. Given a sample of size n, where each individual is chosen to be in treatment ($W = 1$) or control ($W = 0$) with equal probability $\frac{1}{2}$, let $\widehat{\tau}_n, \widehat{\alpha}_n, \widehat{\beta}_n$ be the empirical squared error minimizers. Give the limiting distribution of $\widehat{\tau}_n$. That is, give the value of the asymptotic variance $\sigma^2(\tau)$ in the limiting normal

$$\sqrt{n}(\widehat{\tau}_n - \tau^{\text{mse}}) \xrightarrow{\text{dist}} \mathsf{N}\left(0, \sigma^2(\tau)\right).$$

Question 2.3 (Constructions of conformal confidence sets, 15 points)**:** Suppose we have set-valued mappings $C_\tau : \mathcal{X} \rightrightarrows \mathcal{Y}$, meaning that $C_\tau(x) \subset \mathcal{Y}$, indexed by $\tau \in \mathbb{R}_+$, where

$$C_\tau(x) \subset C_{\tau+\delta}(x)$$

[1] We have $\mathbb{E}[v^T x(u^T x)^2] = \sum_{i=1}^d v_i\mathbb{E}[x_i(u^T x)^2]$, and (w.l.o.g. taking $i = 1$) we observe $\mathbb{E}[x_1(u^T x)^2] = \sum_{i,j=1}^d u_i u_j\mathbb{E}[x_1 x_i x_j]$. Then note that $\mathbb{E}[x_1 x_i x_j] = 0$ for any coordinates i, j when $x \sim \mathsf{N}(0, I)$.

for all $\delta \geq 0$, where $\lim_{\tau \to \infty} C_\tau(x) = \mathcal{Y}$ (that is, for large enough τ the confidence set $C_\tau(x)$ includes all of $\mathcal{Y}$). Define

$$s(x, y) := \inf \{\tau \in \mathbb{R} \mid y \in C_\tau(x)\}. \tag{2.1}$$

You are given a sample $(X_i, Y_i)_{i=1}^n \overset{\text{iid}}{\sim} P$ of size n and define $S_i = s(X_i, Y_i)$ for each i, then set

$$\widehat{\tau}_n := \text{the } (1 + 1/n)(1 - \alpha) \text{ quantile of } \{S_i\}_{i=1}^n.$$

Let $\widehat{C} = C_{\widehat{\tau}_n}$ be the associated confidence set.

(a) Using the results from class, show that $\widehat{C}$ is a valid $(1 - \alpha)$ prediction set, that is, on a new example (X_{n+1}, Y_{n+1}) from P,

$$P(Y_{n+1} \in \widehat{C}(X_{n+1})) \geq 1 - \alpha.$$

We now explore different constructions of such confidence sets. Each of these will leverage an already constructed predictor f taking inputs in $\mathcal{X}$.

(b) Let ℓ be a loss function and $\ell(f(x), y)$ be the loss for predicting $f(x)$ on response y. Set

$$C_\tau(x) = \{y \in \mathcal{Y} \mid \ell(f(x), y) \leq \tau\}.$$

Give the value $s(x, y)$ the definition (2.1) yields.

(c) For binary logistic regression, $f : \mathcal{X} \to \mathbb{R}$, $y \in \{\pm 1\}$, and $\ell(f(x), y) = \log(1 + e^{-yf(x)})$. Give the value $s(x, y)$ the definition (2.1) yields for the confidence set in part (b). For a given threshold τ, when is $C_\tau(x)$ a singleton?

(d) For k-class logistic regression, $f : \mathcal{X} \to \mathbb{R}^k$, $y \in \{1, \ldots, k\}$, and $\ell(f(x), y) = \log(1 + \sum_{l=1}^k e^{f_l(x) - f_y(x)})$. Give the value $s(x, y)$ the definition (2.1) yields for the confidence set in part (b). For a given threshold τ, when is $C_\tau(x)$ a singleton?

(e) Let $\mathcal{Y} = \mathbb{R}$ (so we have real-valued responses as in regression), and let $l, u : \mathcal{X} \to \mathbb{R}$ model lower and upper quantiles of Y given X, respectively. (That is, we wish to have $Y \in [l(x), u(x)]$ with a given probability.) Let

$$C_\tau(x) = [l(x) - \tau, u(x) + \tau]$$

where $C_\tau(x) = \emptyset$ if $l(x) - \tau > u(x) + \tau$, i.e., the lower end of the interval is greater than the upper. Give the value $s(x, y)$ the definition (2.1) yields for this confidence set.

Question 2.4 (Conformal sets, 20 points)**:** Consider the following heteroskedastic linear model:

$$y = x^T \beta^\star + \|x\|_2 \varepsilon, \quad \varepsilon \overset{\text{iid}}{\sim} \mathsf{N}(0, 1), \quad x \overset{\text{iid}}{\sim} \begin{cases} \mathsf{N}(e_1, I_d) & \text{w.p. } \frac{1}{2} \\ \mathsf{N}(-e_1, 3I_d) & \text{w.p. } \frac{1}{2} \end{cases} \tag{2.2}$$

where e_1 is the first standard basis vector. Given a sample of size n from this model, let $X \in \mathbb{R}^{n \times d}$ and $Y \in \mathbb{R}^n$ be the usual design matrix and responses. Consider fitting the following two models to this data: first, the standard linear regression model

$$\widehat{\beta}^{\text{mse}} := \operatorname*{argmin}_\beta \|X\beta - Y\|_2^2.$$

This gives predictor $\widehat{f}(x) = x^T\widehat{\beta}^{\text{mse}}$. Second, a quantile model, which attempts to predict the lower and upper α quantiles of the responses y_i. To do this, define the feature mapping $\phi(x) = (x, \|x\|_2) \in \mathbb{R}^{d+1}$, the quantile loss function

$$\ell_\alpha(t, y) := \alpha\,(y - t)_+ + (1 - \alpha)\,(t - y)_+\,,$$

and then find the $(d+1)$-dimensional vectors $\widehat{\theta}_\alpha$ and $\widehat{\theta}_{1-\alpha}$ solving

$$\widehat{\theta}_\alpha = \operatorname*{argmin}_\theta \sum_{i=1}^n \ell_\alpha(\theta^T\phi(x_i), y_i) \;\text{ and }\; \widehat{\theta}_{1-\alpha} = \operatorname*{argmin}_\theta \sum_{i=1}^n \ell_{1-\alpha}(\theta^T\phi(x_i), y_i).$$

This gives lower and upper predictors $\widehat{l}(x) = \phi(x)^T\widehat{\theta}_\alpha$ and $\widehat{u}(x) = \phi(x)^T\widehat{\theta}_{1-\alpha}$.

(a) What is the population counterpart of $\widehat{\beta}^{\text{mse}}$? That is, give $\beta^\star = \operatorname{argmin}_\beta \mathbb{E}[(y - x^T\beta)^2]$.

(b) What are the population counterparts of $\widehat{\theta}_\alpha$ and $\widehat{\theta}_{1-\alpha}$? That is, give

$$\theta^\star_\alpha = \operatorname*{argmin}_\theta \mathbb{E}[\ell_\alpha(\theta^T\phi(x), y)] \;\text{ and }\; \theta^\star_{1-\alpha} = \operatorname*{argmin}_\theta \mathbb{E}[\ell_{1-\alpha}(\theta^T\phi(x), y)].$$

(c) Generate three datasets from the model (2.2), each in dimension $d = 5$ with sample size $n = 400$: a training set, a validation set, and a test set, where $\beta^\star \sim \mathsf{Uni}(\mathbb{S}^{d-1})$. Now, fit the linear predictor $\widehat{f}$ and lower/upper predictors $\widehat{l}, \widehat{u}$ on the training data. Consider the confidence sets

$$\widehat{C}^{\text{mse}}_\tau(x) := \left[\widehat{f}(x) - \tau, \widehat{f}(x) + \tau\right] \;\text{ and }\; \widehat{C}^{\text{q}}_\tau := \left[\widehat{l}(x) - \tau, \widehat{u}(x) + \tau\right].$$

Using the validation data, use conformal inference to choose τ so that $\mathbb{P}(Y^* \in \widehat{C}_\tau(X^*)) \geq 1 - 2\alpha$ for each of these confidence sets, where $\alpha = .025$. Repeat this experiment $T = 100$ times and give the (empirical) coverage you obtain on the test set for each method.

(d) As in part (c) (with $T = 100$ experiments), give average the empirical coverage on the following two subsets of the test set:

$$S_{\text{left}} := \{i \mid e_1^T x_i \leq 0\} \;\text{ and }\; S_{\text{right}} := \{i \mid e_1^T x_i > 0\}.$$

Explain your result in a few words.

Question 2.5 (A loan data analysis challenge, 40 points)**:** A company in Chile uses crowdsourcing to fund loans to the public, as a means to offer relief from the high bank interest rates. The data in this challenge consists of historical loan records for a sample of 9000 past customers. The variables characterize some aspects of the loan, such as duration, amount, interest rate and many other more technical features of the loans. There are also a number of qualitative variables, such as reason for loan, quality rating of the borrower and others. The response variable y of interest is `default`: a 0-1 variable indicating whether or not the borrower has defaulted on their loan payments.

The company would like to build a default risk score so that they can target high-risk customers early and perhaps preempt the default event, which ends up costly for all involved. (The fraction of defaults in the entire population is around 7%.) The training data `loan-train.csv` represents a sample of 3000 defaulters, and 6000 non-defaulters, and contains 30 features and the binary outcome `default` (in the first column). The file `loan-testx.csv` consists of a random sample of 10000 other customers from the general pool. For these you are provided only the 30 features.

Your job is to build a *risk score*, that is, a model that estimates the probability of default $y = 1$. Feel free to use any of the tools discussed in the lectures of this class (or beyond). Some packages that may be useful include `pytorch`, `xgboost`, and just regular old logistic regression. Describe what you implemented, how you selected your final model. The only thing you need to submit is a text file with 10000 lines; on each line, your should have your predicted risk estimate for each test customer, in the same order as `loan-testx.csv`. Submit this as a .txt file on Gradescope for the online portion of HW2.